The EMBO Journal vol.5 no.4 pp.819–822, 1986

Using known substructures in protein model building and crystallography

T.Alwyn Jones and Sören Thirup

Department of Molecular Biology, Box 590, Biomedical Center, S-751 24 Uppsala, Sweden

Communicated by M.Levitt

Retinol binding protein can be constructed from a small number of large substructures taken from three unrelated proteins. The known structures are treated as a knowledge base from which one extracts information to be used in molecular modelling when lacking true atomic resolution. This includes the interpretation of electron density maps and modelling homologous proteins. Models can be built into maps more accurately and more quickly. This requires the use of a skeleton representation for the electron density which improves the determination of the initial chain tracing. Fragment-matching can be used to bridge gaps for inserted residues when modelling homologous proteins.

Key words: protein modelling/retinol binding protein/structure prediction

Introduction

Single-crystal X-ray diffraction is still the most powerful method for obtaining the three-dimensional structure of a protein molecule. Although many technical improvements have been made since the structure of myoglobin was determined by Kendrew *et al.* (1960), the interpretation of the experimentally determined electron density map remains a difficult point in the process. For many reasons these maps are rarely of sufficient quality to show individual atoms and will often contain breaks in main chain density.

Map interpretation has been made easier by the fact that proteins contain significant amounts of regular structure such as α-helices and β-strands, predicted by Pauling and co-workers (Pauling and Corey, 1951; Pauling *et al.*, 1951), and standard turns (Venkatachalam, 1968; see Richardson, 1981, for an extensive review). A model with the correct conformation can then be made and fitted to even a relatively poor piece of density. Such a model may be made of wire (Richards, 1968) or generated in a computer graphics system (Jones, 1982).

This building-block approach to protein modelling can be expanded to include all fragments making up the molecule. We show that a protein can be constructed from large fragments of just a few proteins. Such substructures can also be easily matched to a suitable representation of the electron density, e.g. to the skeletonized density of Greer (1974). Because many errors and ambiguities exist in such a skeleton, extensive adjustments may be required and we have therefore implemented these methods in the interactive graphics program, FRODO (Jones, 1978, 1982, 1985).

Although we emphasize the use of fragment-fitting in protein crystallography, the technique is useful in a number of modelling activities involving non-atomic resolution data. This includes modelling homologous proteins where it can suggest a number of possible conformations, and in n.m.r. spectroscopy where fragments can be located to satisfy local interatomic distance measurement (Kraulis and Jones, in preparation).

Results and Discussion

Searching for fragments of similar structure

The main domain of retinol binding protein (RBP) consists of an unusual eight-stranded up-and-down β-barrel that encapsulates the retinol molecule (Newcomer *et al.*, 1984). There are seven reverse turns between these β-strands, two of which, residues 48–52 and 124–128, have a similar but unusual main chain hydrogen bonding scheme.

In this scheme carbonyl oxygen O_i forms a hydrogen bond with peptide nitrogen N_{i+3}, and N_i forms one with O_{i+4}. In both turns residue i+3 is a glycine, and the main chain torsion angles correspond to a type I turn (Venkatachalam, 1968) followed by a bulge (Richardson *et al.*, 1978). We found that the relevant Cα atoms of 48–52 can be matched to 124–128 with a root mean square (r.m.s.) deviation of only 0.23 Å (Figure 1). This turn conformation had not been identified as a standard substructure in the review by Richardson (1981) but its frequency in RBP suggested that it may be a common template in anti-parallel strands. To test this hypothesis, we searched the entire protein data bank (Bernstein *et al.*, 1977) and found this substructure

Table I. Results of building RBP from three other proteins

RBP residue	Matching protein	Protein residue number	R.m.s. deviations (Å)
4–11	HCAC	30	0.64
12–17	ADH	77	1.02
18–21	HCAC	130	0.04
22–33	HCAC	204	1.28
34–39	ADH	9	0.39
40–47	STNV	159	0.67
48–52	ADH	123	0.29
53–62	STNV	132	0.56
63–67	ADH	308	0.34
68–78	STNV	69	1.04
79–86	HCAC	94	1.02
87–92	ADH	312	0.24
93–97	ADH	110	0.26
98–106	ADH	22	0.91
107–114	STNV	122	1.28
114–121	STNV	61	0.54
122–128	ADH	121	0.66
129–139	STNV	56	1.03
140–144	STNV	28	0.31
145–161	ADH	323	0.79
161–167	HCAC	133	0.80
168–173	HCAC	56	0.53

The matching protein residue number is the internal residue count of the first residue in the matching zone. As such, it is not necessarily the same as the residue name. The r.m.s. deviation is the result of a least squares fit.

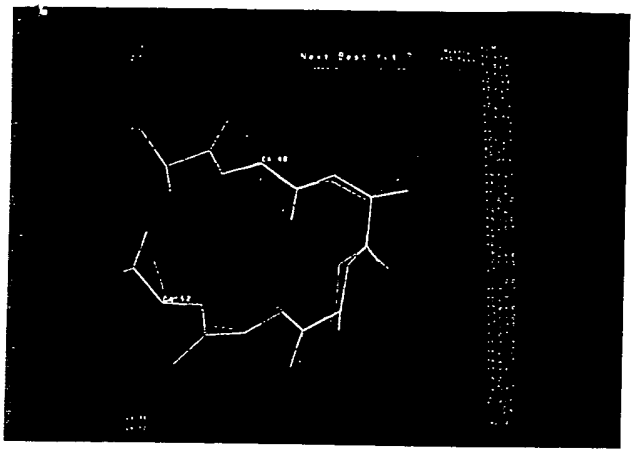

Fig. 1. Matching the reverse turn near RBP residue 50. The backbone atoms of residues 48–52 are shown in an atom colouring convention (carbons are yellow, nitrogens are blue and oxygens are red). The green fragment is RBP residues 124–128 and matches with a deviation of 0.23 Å.

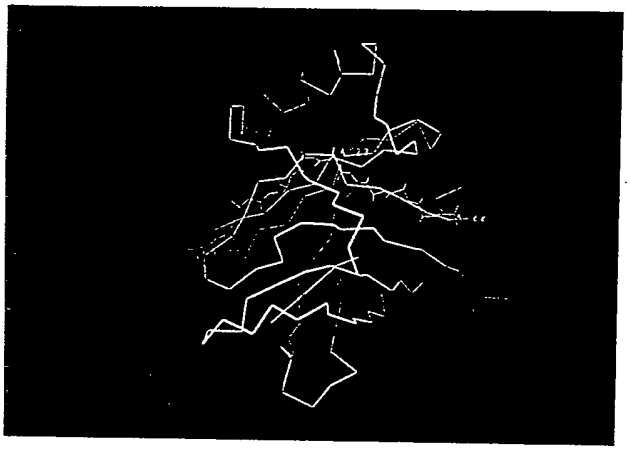

Fig. 2. Searching for a match to all of strand A in RBP, including the bulge at residue 27. The green fragment is from carbonic anhydrase *C*.

n 23 different proteins using an r.m.s. cut-off of 0.5 Å for mat-hing all main chain atoms. The same substructure has been ecently identified by Sibanda and Thornton (1985) from an ex-ensive study of turns between anti-parallel strands.

The ease with which we found an unknown conformation rompted the question of whether, indeed, any part of RBP was nique. The matching fragments shown in Table I were obtain-d by trial and error at the display from the refined coordinates f only three proteins: satellite tobacco necrosis virus, STNV Jones and Liljas, 1984b), apo-alcohol dehydrogenase, ADH Jones and Eklund, in preparation) and human carbonic anhydrase , HCAC (T.A.Jones, E.Eriksson and A.Liljas, in preparation). f a fragment matched, the region was extended and the match epeated. A large substructure is shown in Figure 2. After ebuilding the whole molecule, it was regularized to remove the iscontinuities that had been introduced between the fragments. he main chain r.m.s. deviation to the starting model was 1.0 Å. Considering we used fragments from only three proteins and that e located and combined them in a very simple way, this is a rprising result both in terms of the goodness-of-fit and in the umber of fragments used.

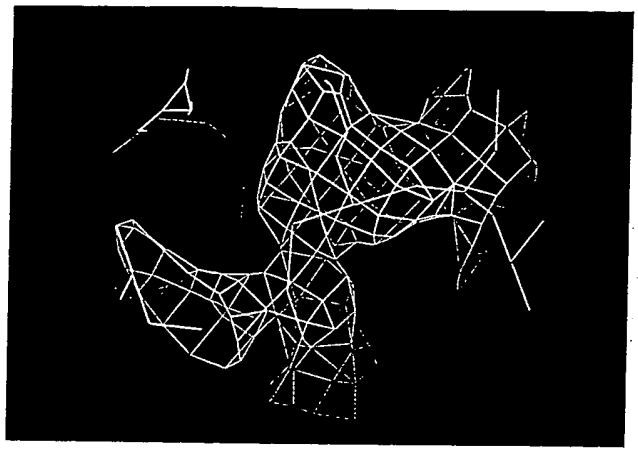

Fig. 3. Portion of the 3.1 Å electron density map of RBP with a calculated skeleton. The skeleton atoms have been automatically classified as main chain (lilac) and side chain (green).

A more elegant method of finding the best set of fragments has been suggested (M.Levitt, private communication) that uses a dynamic programming algorithm similar to that employed in sequence comparisons (Needlemann and Wunsch, 1970). This finds the minimum number of fragments required to build the structure where each fragment matches the structure to within a pre-set limit. With the criterion that Cαs are matched to within 1 Å r.m.s., this method builds RBP from 15 fragments, and with a 0.5 Å cut-off it requires 20 fragments.

Protein crystallographic applications

The construction of an initial model from an electron density map is frequently a difficult task for even highly trained scientists. The process is usually complicated by lack of resolution in the X-ray data, lack of isomorphism in the heavy atom derivatives and sometimes by lack of an amino acid sequence. The crystallographer is faced with long range problems such as getting the correct chain tracing, and local problems such as the correct orientation of peptide planes. Jones (1982) showed that even simple peptide plane errors could not be automatically removed by refinement programs. If the rest of the model is sufficiently accurate, maps calculated with phases obtained from the model coordinates can be inspected to locate and correct errors (Jones, 1982). More usually in crystallographic refinement, the model (and hence the phases) gradually improves but requires many cycles of model refitting and refinement (Remington *et al.*, 1982). There are even reports that removing incorrect parts of the structure from the phases can still leave a ghost of the incorrect structure in the map (Finzel *et al.*, 1984). It is therefore of great benefit to start with as accurate a structure as possible.

Various methods have been used to build models into maps with computer graphics. Our experience with the program FRODO (Jones, 1978, 1982, 1985) suggests that it is first necessary to determine the protein fold from contoured mini-maps drawn on plastic sheets. If secondary structure elements are recognized they can be constructed and fitted as rigid groups to a set of rough atomic guide points. Any gaps can be filled in based on a few guide points per residue.This produces a rough starting model whose main chain follows the trace determined on the mini-map. It then requires refitting to closely match the density, possibly using automated techniques (Jones and Liljas, 1984a).

A different approach has been suggested by Greer (1974) that

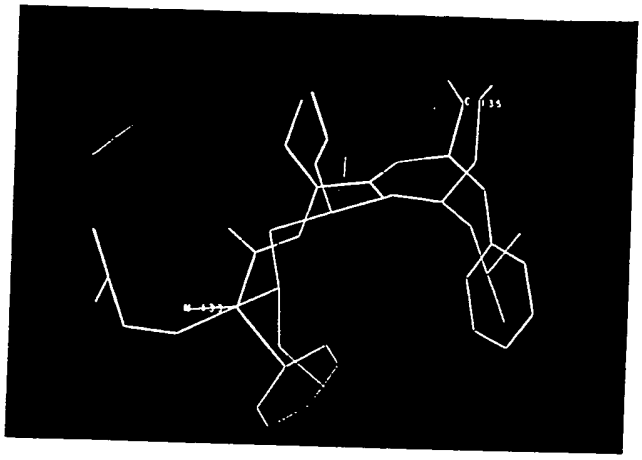

Fig. 4. The skeleton of Figure 3 has been re-defined to suggest a different chain tracing. In this hypothesis, the orange line represents the new main chain. A number of bonds have been made and broken to satisfy this hypothesis. The final refined coordinates (Tyr−Ser−Phe) are drawn with a standard atomic colouring scheme.

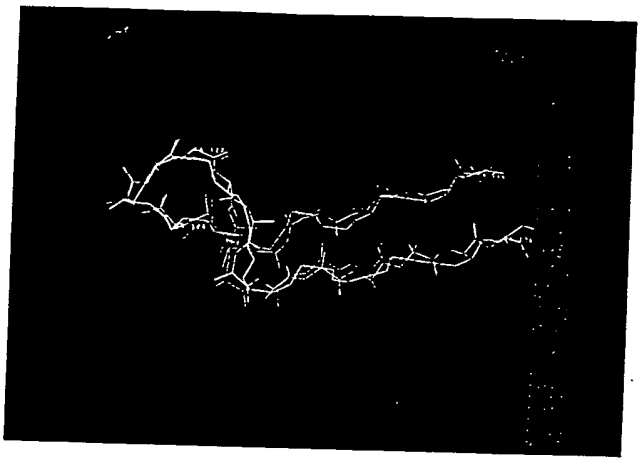

Fig. 5. The main chain skeleton (in orange) has been fragment-fitted (in green) five residues at a time and with one residue overlap. The final crystallographically refined RBP residues 116−136 are drawn with the atom colouring scheme.

attempts to automate model building. The first step in this procedure reduces the electron density map to a set of connected points that follow the density. This skeletal representation makes it much easier to recognize branch points that may represent side chains or, at higher resolution, carbonyl groups. The method was tested on good quality 2 Å and 3 Å maps of known structures (Greer, 1974, 1976) and could produce provisional main chain coordinates. The many errors usually present in skeleton connectivity probably explain why the method has not been widely used.

A different skeletonization algorithm using critical point networks (Johnson, 1978) has been combined with computer graphics by Pique (1984) and co-workers.

A part of the 3.1 Å multiple isomorphous replacement map of RBP is shown in Figure 3. This shows the usual electron density contours with a Greer skeleton calculated from the map. The skeleton colouring shows a calculated main chain/side chain assignment made according to the lengths of connected pieces. This region corresponds to the tripeptide 133−135 (Tyr−Ser−Phe) and was the starting point of our initial map interpretation. It illustrates the problems associated with low resolution maps: the main chain skeleton shows a break due to a local narrowing of the density, and a pair of hydrogen bonding side chains (the serine and a tyrosine that is not drawn) form a continuous density that is assigned main chain status. Figure 4 shows the same region after interactively locating and correcting these errors, and defining the main chain skeleton as an acceptable trace. Figures 3 and 4 show the important use of colour to illustrate the current skeleton assignments.

Fragments from known structures can be matched to the skeleton. This first requires positioning putative Cα atoms along the skeleton and can be done automatically (with restraints that neighbouring Cα atoms are ~3.8 Å apart) or by explicitly defining a skeleton point to be a Cα atom. To ensure a good and quick match, one usually fits a length of skeleton corresponding to a fragment of 5−7 residues. Adjacent fragments are often overlapped by one residue because the carbonyl group of the last residue in a fragment plays no role in the matching.

Fragment-fitting to the skeleton, over the region corresponding to RBP residues 116−136 (Figure 5), gives a model with an r.m.s. deviation to the final refined coordinates of 0.95 Å

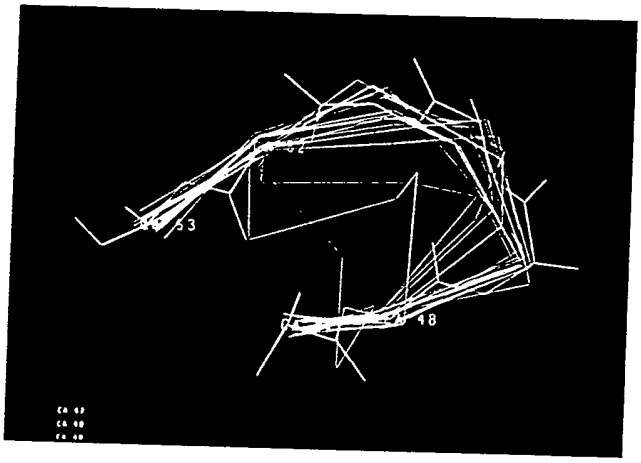

Fig. 6. In the RBP loop 47−53, residues 49−51 have been excluded from the fragment search. The coloured traces are the 20 best fits to the remaining four residues. The correct RBP chain is a member of the dominant cluster of 14 traces.

for all main chain atoms. Our original MIR model was the result of many hours of careful fitting, but has a significantly higher r.m.s. deviation of 1.30 Å to the final coordinates.

The skeleton has the added advantage of giving an overall view of the density for which one previously relied on mini-maps. However, it is a great improvement since it can be easily changed, saved, restored and viewed from any direction.

Model building homologous structures

The number of known protein sequences far exceeds the number of known structures. Thus, for each newly determined structure there is usually at least one other protein with some sequence homology. For example, the *Escherichia coli* DNA polymerase 1 Klenow fragment (Ollis *et al.*, 1985a) could immediately be used to model T7 DNA polymerase (Ollis *et al.*, 1985b). Model building homologous proteins relies on defining structurally conserved and variable regions (Greer, 1981). Amino acid mutations in conserved regions are easily carried out with programs such as FRODO. Insertions and deletions in the variable regions are much more difficult to model. In these regions we are fre-

quently faced with the problem: how does one go between two points in space using a certain number of amino acids?

Substructure matching is able to provide some answers to this question. By way of illustration, we shall again refer to loop 47–53 in RBP. When a search is made for this loop among the coordinates of 37 highly refined proteins, all of the 20 best matches have the RBP conformation. Excluding Gly 51 from the search also gives 20 essentially identical traces. Excluding residues 49–51 gives the set of Cα traces shown in Figure 6. Fourteen of the 20 traces are similar to the loop observed in RBP and of these, seven had a glycine equivalent to Gly 51. A second cluster of three conformations is also apparent. We are currently investigating various length loops and deletions to more accurately determine the probability of identifying the correct substructure.

Conclusions

Our initial experiments suggest that proteins can be constructed from large building blocks whose exact size and number remain to be determined.

We have extracted from the protein data bank the best refined sets of coordinates to use as a knowledge base for structure analysis. A fast search and matching algorithm allows one to interactively model substructures from this database under conditions made difficult by a lack of high resolution data.

Our computer graphics implementation of density skeletonization gives an improved overview of a possible chain trace. It also contains sufficient detail to build a model with fragment-fitting which is at least as good as can be obtained by careful manual fitting. The speed with which we can build a model from a skeleton makes it much easier to test chain tracing hypotheses.

Materials and methods

Diagonal plot algorithm for locating similar conformations

Efficient techniques have been developed to find the best least squares fit of one set of points to another set (Kabsch, 1978; McLachlan, 1979). The goodness-of-fit can then be judged by the r.m.s. deviation between one set of points and the correctly transformed second set. Alternative methods can be formulated; in particular we have used the interatomic diagonal plot (Phillips, 1970). This consists of a matrix of distances where element (i,j) is the distance between points i and j. When these points represent protein Cα atoms, the plot can be used to recognize domains and structural motifs (Rossmann and Liljas, 1974).

If two fragments have the same structure, they will also have the same set of inter-Cα distances. However, the reverse is not true. Our distance matching algorithm is 35 times faster than a least squares algorithm when comparing five points. It is therefore used as a sieve to locate similarities which are then tested with the least squares algorithm. The goodness-of-fit of each fragment is judged by the sum:

$$\Sigma(d_m - d_n)^2$$

where d_m is an inter-Cα distance in one structure and d_n is the equivalent distance in the second structure, and the sum is taken over the relevant distances.

The protein Cα distances are pre-calculated by a Fortran program that accepts all commonly used coordinate files. A fragment of five residues can be searched in a library of 34 proteins (containing 5271 residues) in ~3 s on a Vax 750 computer.

Electron density skeletonization

This is a two stage procedure. The first step creates a set of linked points from an electron density map using essentially Greer's algorithm (Greer, 1974). This first removes all points below a pre-set value. Multiple passes are then made through the map with an increasing threshold. A point will not be removed if a hole is created, or if it is a tip or single point. All points with a value equal to the current threshold will then be removed unless they are needed to preserve continuity. This algorithm results in a connected trace of points which is sensitive to the starting base value. We find that contoured electron density is best viewed at one standard deviation, while the skeleton is best calculated with a base level and increment of ~1.3 and 1.0 SD, respectively.

In the second stage each 'atom' in the skeleton is given a status defining it as part of the main chain or of a side chain. This is done according to the length of the linked list containing the atom. The program also provides extra unconnected atoms that may be used later.

Both programs are written in Fortran, and skeletonize a map of 56 × 49 × 77 points in 14 min on a Vax 750.

Graphics interface

We have implemented our FRODO enhancements on a coloured line drawing Evans and Sutherland PS330. The calculated skeleton can be changed by moving its atoms, by re-defining connectivity, and by re-assigning the atomic status. A third status is available which we normally use to define our currently accepted main chain trace. Colour is vital to show the current skeleton assignments (Figures 3 and 4).

Two fragment matching options are available. One places Cα atoms along a linked list of skeleton or protein atoms such that each is positioned ~3.8 Å from its neighbour, unless forced to accept particular points as Cαs. In the second option, one explicitly defines which Cα atoms in a protein fragment are to be used to make a match.

The Cα traces of the 20 best matches can be viewed (Figure 6) and each can be seen in turn as a stripped poly-alanine chain (Figure 1). The coordinates of any of these fits can be incorporated into the FRODO atomic data set. The fit of each residue can then be further improved either manually (Jones, 1978), automatically by real space methods (Jones and Liljas, 1984a) or with a new option that matches all of the residue, including the side chain, to the skeleton. Any combination of accepted main chain, side chain or automatically assigned main chain atoms can be viewed. This gives one the flexibility to view details or to get a large volume overview.

Acknowledgements

T.A.J. acknowledges a research position from the Swedish Natural Science Research Council and S.T. acknowledges an EMBO long fellowship. These methods have been added to the PS300 implementation of FRODO. We thank F.Guiocho, J.Pflugrath, M.Saper and J.Sack for making it available to use. We also wish to thank M.Levitt for his encouragement and interest in this work. Terese Bergfors prepared the manuscript with her usual efficiency.

References

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Finzel,B.C., Poulos,T.L. and Kraut,J. (1984) *J. Biol. Chem.*, **259**, 13027–13036.
Greer,J. (1974) *J. Mol. Biol.*, **82**, 279–302.
Greer,J. (1976) *J. Mol. Biol.*, **104**, 371–386.
Greer,J. (1981) *J. Mol. Biol.*, **153**, 1027–1042.
Johnson,C.K. (1978) *Acta Crystallogr., A.*, **34**, S353 (abstract only).
Jones,T.A. (1978) *J. Appl. Crystallogr.*, **11**, 268–272.
Jones,T.A. (1982). In Sayre,D. (ed.), *Computational Crystallography*. Clarendon Press, Oxford, pp. 303–317.
Jones,T.A. and Liljas,L. (1984a) *Acta Crystallogr., A.*, **40**, 50–57.
Jones,T.A. and Liljas,L. (1984b) *J. Mol. Biol.*, **177**, 735–767.
Jones,T.A. (1985) *Methods Enzymol.*, **115**, 157–171.
Kabsch,W. (1978) *Acta Crystallogr., A.*, **34**, 827–828.
Kendrew,J.C., Dickerson,R.E., Stanberg,B.E., Hart,R.G., Davies,D.R., Phillips,D.C. and Shore,V.C. (1960) *Nature*, **185**, 422–427.
McLachlan,A.D. (1979) *J. Mol. Biol.*, **128**, 49–79.
Needlemann,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
Newcomer,M.E., Jones,T.A., Aqvist,J., Sundelin,J., Eriksson,U., Rask,L. and Peterson,P.A. (1984) *EMBO J.*, **3**, 1451–1454.
Ollis,D.L., Brick,P., Hamlin,R., Xuong,N.G. and Steitz,T.A. (1985a) *Nature*, **313**, 762–766.
Ollis,D.L., Kline,C. and Steitz,T.A. (1985b) *Nature*, **313**, 818–819.
Pauling,L. and Corey,R.B. (1951) *Proc. Natl. Acad. Sci. USA*, **37**, 729–740.
Pauling,L., Corey,R.B. and Branson,H.R. (1951) *Proc. Natl. Acad. Sci. USA*, **37**, 205–211.
Pique,M. (1984) *J. Mol. Graphics*, **2**, 59 (abstract only).
Phillips,D.C. (1970). In Goodwin,T.W. (ed.), *British Biochemistry, Past and Present*. Academic Press, London, pp. 11–28.
Remington,S., Weigland,G. and Huber,R. (1982) *J. Mol. Biol.*, **158**, 111–152.
Richards,F.M. (1968) *J. Mol. Biol.*, **37**, 225–230.
Richardson,J.S., Getzoff,E.D. and Richardson,D.C. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 2574–2578.
Richardson,J.S. (1981) *Adv. Protein Chem.*, **34**, 168–339.
Rossmann,M.G. and Liljas,A. (1974) *J. Mol. Biol.*, **85**, 177–181.
Sibanda,B.L. and Thornton,J.M. (1985) *Nature*, **316**, 170–174.
Venkatachalam,C.M. (1968) *Biopolymers*, **6**, 1425–1436.

Received on 2 January 1986